**Impact of restaurant smoking restrictions on smoking rate**

**In this problem set, you are going to use SMOKE.DTA data. In this data set, you have information on individuals’ smoking behaviour and some other individual and locational characteristics for a random sample of single adults from the United States.**

**1) What is the share of people in the sample who smokes (you need to generate a new binary variable indicating whether the person is ”smoking”)? What is the share of people who resides in a state with restaurant smoking restrictions? What is the difference in average smoking probability between states with restaurant smoking restrictions and states without restrictions?**

*share of smokers = 0.3841388*

*share of people who lives in a smoke-restricted state = 0.2465923*

After generating the “smoking” variable, the share of people in the sample who smokes is observed as 38%. The share of people that resides in a state with restaurant smoke restrictions is 25%. If we observe restricted and non-restricted states seperately, in states with restaurant smoking restrictions the share of smokers is 31% and in states without restaurant smoking restrictions the share is 41%. The difference in non-restricted states and restricted ones is approximately 10 percentage points.

**2) Using the information available in the data set, estimate a linear probability model that examines the determinants of smoking probability (note that the dependent variable will be a binary variable). Include all potentially relevant variables that might affect smoking behavior (you need to decide which variables might affect smoking behavior). Interpret the signs and magnitudes of the coefficients that are significant at 10% significance level.**

The linear regression model

with binary dependent variable which is here, is called the linear probability model.

To estimate a linear probability model that examines the determinants of smoking probability, we need to use “smoking” variable that we created in Q1 as binary dependent variable. The potential relevant variables that might affect smoking behavior are educ (years of schooling), cigpric (price of cigarette), age, income, restaurn (restaurant smoking-restricted states) and white (=1 if white).

|  |  |
| --- | --- |
|  | (1) |
| VARIABLES | smoking |
|  |  |
| educ | -0.0249\*\*\* |
|  | (0.00594) |
| age | -0.00364\*\*\* |
|  | (0.00101) |
| income | 2.66e-06 |
|  | (1.97e-06) |
| cigpric | -0.000889 |
|  | (0.00360) |
| restaurn | -0.104\*\*\* |
|  | (0.0399) |
| white | -0.0125 |
|  | (0.0519) |
| Constant | 0.883\*\*\* |
|  | (0.237) |
|  |  |
| Observations | 807 |
| R-squared | 0.040 |

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

By looking the table above, significant variables at 10% significance level are educ, age, restaurn.

The interpretation of this linear probability model is as follows;

* One year of additional education decreases the possibility of smoking by 2.5 percentage points.
* If age increases by one year, smoking possibility decreases by 0.3 percentage points.
* In states with restaurant smoking restrictions the probability of smoking is 10.35 percentage points less than states without restaurant smoking restrictions.

**3) Estimating a new specification, check whether the effect of age on smoking probability is quadratic or not (assuming that you included ”age” in question 2 in linear form). At what age does the impact of age on smoking probability becomes negative?**

Comparing with the model in Q2, age variable was significant and in this model age, agesq both are significant variables with p-value 0. Additionally, R-squared is increased from 0.0401 to 0.0618, our models explanatory power is greater than the previous one. Based on these observations, we can conclude that the effect of age on smoking probability is quadratic.

To answer the question “At what age does the impact of age on smoking probability becomes negative?” we need to solve this equation (derived from the model).

At the 76 (closing 77), impact of age on smoking probability becomes negative.

**4) Now, estimate your model including the income and cigarette prices in logarithmic form instead of level form (assuming that you included them in question 2 in level form). Interpret the signs and magnitudes of the estimated coefficients for these two variables.**

In these version of the model, using logarithmic form for income and cigpric helps us to make better estimations. To interpret, one percent increase of income increases smoking probability by 0.012 percentage points and one percent increase of cigarette prices decreases the probability of smoking by 0.069 percentage points.

**5) You might think that the impact of restaurant smoking restrictions on smoking probability might be different for white and non-white individuals. Using an interaction variable approach, test whether this hypothesis is true at 10% significance level.**

To understand and interpret the interaction between white and restaurn variables, we should use interaction variable approach. Adding an interaction term to this model can clarify the correctness of our hypothesis.

This restwhite coefficient (restaurn\*white) is not significant at 10% significant level since p-value is 0.321 > 0.05 (two sided 0.10 / 2) and t-statistic is 0.99 < 1.65. Thus, our hypothesis that claims different impact of restrictions to white and non-white people, is not true at 10% significance level. Because of restwhite being insignificant, it can be dropped (restwhite = 0) and in this case fails.

**6) Can we consider the coefficient for restaurn as the causal effect of restaurant smoking restrictions on individuals’ smoking probability? Discuss whether there might be an endogeneity problem here. Provide an example of omitted state specific factor that could lead to a bias in the** **estimated effect of smoking restrictions and discuss the direction of the potential bias that might arise because of this omitted factor.**

When we analyze the models from the previous questions, the effect of restaurant smoking restrictions on individuals’ smoking probability is huge (about 10 percentage points) and it remains almost constant despite the changes we made in each question. Although it seems that restaurn has a high causal effect on smoking probability, it can not be proven just by these regression models. In order to deduce causality, the assumption of exogeneity must hold. Exogeneity assumes the changing our target variable and holding other factors constant (independent from each other) is possible.

But in this situation we might conclude following statement, “In the states with restaurant smoking restrictions, people and government are so awared against smoking and its unhealthy effects that they decide to legislate smoking restrictions for restaurats.” with other words, smoking restrictions might be the result itself instead of being a reason for smaller share of smokers. For example if we hold the propensity to consume cigarettes or a state level anti-smoking sentiment in a variable, this omitted state specific factor causes a potential bias. When anti-smoking sentiment is omitted (in our model), restaurn will be underestimated because all the negative effects of the anti-smoke attitudes will be added to restaurn variable. The direction of the potential bias is negative.

**Impact of Job Training Grant**

**Use the data from JTRAIN.DTA for this exercise.**

**7) Consider the simple regression model ( log(scrap) = β0 + β1grant + u ) where scrap is the firm scrap rate and grant is a dummy variable indicating whether a firm received a job training grant. Can you think of some reasons why the unobserved factors in u might be correlated with grant? Provide example.**

Due to some significant omitted variables in the model (that have an impact on both grant and scrap but not included in the regression model) with infinitely many other potential reasons, the unobserved factors in “u” might be correlated with grant. For example, education and qualification of the workers, critical attributes of the workforce can make the firm more likely to receive a training grant. Other firm characteristics can also be the reason for a job training grant such as attrition-rate, complexity of operations etc. To provide detailed answer to this question, understanding the process of granting would be useful. Besides these examples, the grant possibly requires some specific criteria for firms.

**8) Estimate the simple regression model using the data for only 1988 (You should have 54 observations.) Does receiving a job training grant significantly lower a firm’s scrap rate?**

1988 Simple Regression Model:

# of observations = 54

By analyzing the regression model, the conclusion does not support the statement in the question. There is no evidence that supports receiving a job training grant significantly lowers a firm’s scrap rate. In fact, the coefficient of grant is positive (not statistically significant) according to 1988’s data.

**9) Now, add an additional explanatory variable (to the model in question 8) indicating the log scrap rate of the company in year 1987 (lscrap1). Interpret the coefficient on grant. Is it statistically significant at the 10% significance level? How do you explain the change in the coefficient of grant between two models (from question 8 to 9)?**

|  |  |
| --- | --- |
|  | (1) |
| VARIABLES | lscrap |
|  |  |
| grant | -0.254\* |
|  | (0.147) |
| lscrap\_1 | 0.831\*\*\* |
|  | (0.0444) |
| Constant | 0.0212 |
|  | (0.0891) |
|  |  |
| Observations | 54 |
| R-squared | 0.873 |

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

In first model, we can interpret the model as receiving a training grant increases the firm’s scrap rate by 5.66%. After adding lscrap\_1 to the model, if a firm receives a job training grant its scrap rate is expected to decrease by 25.397%. Adding lscrap\_1 (the log scrap rate of the company in year 1987) changed the sign and value of grant and increased the statistical significance of it. T-value of grant equals -1.73 which is smaller than -1.65 (10% significance level) and p-value is almost 0. Hence, the coefficient of grant is statistically significant at 10% significance level.

**Marijuana usage and Wage**

**Suppose you collect data from a survey on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question in the survey is: ”On how many separate occasions last month did you smoke marijuana?”**

**10) Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, ”Smoking marijuana five more times per month is estimated to change wage by x%.”**

Since we want to estimate the percentage change in wage we need to build a log-linear model. The regression model might look like that:

**11) Write a model that would allow you to test whether marijuana usage has different effects on wages for men and women. How would you test that there are no differences in the effects of marijuana usage on wage for men and women?**

To test the effects of marijuana usage on wages gender dependently, we need to use an interaction variable between gender and marijuana variables of previous model. The new model is going to be like that:

To test the gender effect of marijuana’s wage impact, we could test the following hypothesis with a t-test:

First we need to find t-statistic:

Then we should look for the ciritical value based on the (1) percentile in the t dist with n-1 degrees of freedom. If the absolute t-value is greater than the critical value, we should reject the null hypothesis.

**12) Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and 3 heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage by using this categorical variable.**

To avoid perfect collinearity, we can omit the nonuser variable as our reference point. Based on this, our new model would look like this:

It will give us better estimation by using these categorical variables.

**13) Discuss whether it is possible to estimate the causal effect of marijuana usage on wage based on this survey data? What might be the problem in identifying the causal effect here? Provide an example.**

Explaining the causality completely is not possible since we can not have a perfectly explanatory and unbiased data or we can not ensure the necessary exogeneity. There might be some problems about our survey data that can inhibit our efforts to estimate causal effect of marijuana usage on wage. One of them is our survey might be non representative of the population. Self-selection and social desirability bias lead this problem. For example, if you ask an individual about his/her marijuana consumption behavior, response can be manipulated by the fear of looking like an addict. To give an other example, the location or particular time we conducted survey may affect the results as well. Eliminating all these biases and surveying a truly random sample is almost impossible. Thus, there will be always potential problems with drawing causal inference using the survey like this.

**Mert Ekici 26772**